8th Vilnius Prob. Conf.
B. Grigelionis *et al.* (Eds)
2002 Vilnius

On Bayesian Adaptation

SUBHASHIS GHOSAL

Department of Statistics, University of North Carolina

JYRI LEMBER

Eurandom, Eindhoven

AAD VAN DER VAART
Department of Mathematics, Vrije Universiteit Amsterdam, De Boelelaan 1081, Amsterdam

Abstract. We show that Bayes estimators of an unknown density can adapt to unknown smoothness of the density. We combine prior distributions on each element of a list of log spline density models of different levels of regularity with a prior on the regularity levels to obtain a prior on the union of the models in the list. If the true density of the observations belongs to the model with a given regularity, then the posterior distribution concentrates near this true density at the rate corresponding to this regularity.

Key words: Posterior distribution, rate of convergence, adaptation, sieves, splines, model selection.

Mathematics Subject Classifications (2000): 62G15, 62G20, 62F25.

1. INTRODUCTION AND RESULTS

Consider the problem of estimating a probability density p on the unit interval based on a random sample $X_1, \ldots, X_n$ from this density. If p is known to possess α derivatives, then it is well-known that p can be estimated at the rate $\epsilon_{n,\alpha} = n^{-\alpha/(2\alpha+1)}$, relative to, for instance, the Hellinger or L_2-distance on the set of probability densities. A variety of methods achieve this rate, which is known to be optimal in a minimax sense if nothing more is known concerning p. Furthermore, it is well known that the rate $\epsilon_{n,\alpha}$ can be achieved even if the value of α is not known a-priori. So-called *rate-adaptive* estimators achieve the rate $\epsilon_{n,\alpha}$ if p is α-smooth for a selection of values of α simultaneously. The purpose of this paper is to investigate such rate-adaptation within a fully Bayesian set-up, and whether it can be achieved through Bayesian model

selection.

We investigate this in the setting of log spline density estimation, as first described by Stone (1990). Log spline density models are exponential families constructed as follows. Fix some "order" q, a natural number, throughout. For a given number K partition the half open unit interval $[0,1)$ into K subintervals $\big[(k-1)/K, k/K\big)$ for $k = 1, \ldots, K$. The linear space of splines (with simple knots) of order q relative to this partition is the set of all functions $f : [0,1] \to \mathrm{R}$ that are $q-2$ times differentiable on $[0,1)$ and whose restriction to every of the partioning intervals $\big[(k-1)/K, k/K\big)$ is a polynomial of degree strictly less than q. It can be shown that this is a $J = q + K - 1$-dimensional vector space. A convenient basis is the set of B-splines $B_{J,1}, \ldots, B_{J,J}$, defined e.g. in De Boor (1978). More precisely, let $B_{J,1}, \ldots, B_{J,J}$ be the B-splines of order q for the knot sequence

$$\overbrace{0, 0, \ldots, 0}^{q \text{ times}}, \frac{1}{K}, \frac{2}{K}, \ldots, \frac{K-1}{K}, \overbrace{1, 1, \ldots, 1}^{q \text{ times}},$$

as defined on page 108 of De Boor (1978). The exact nature of these functions does not matter to us here, but their properties are used in a sequence of lemmas in Section 2, which show that various norms of a linear combination of the base elements are comparable to the corresponding norms on the coefficients. This is roughly due to the fact that the basis elements are supported inside intervals of length q/K, which is very small if K is very large relative to q.

For $\theta \in \mathrm{R}^J$ let $\theta^T B_J = \sum_j \theta_j B_{J,j}$ and define

$$p_{J,\theta}(x) = e^{\theta^T B_J(x) - c_J(\theta)}, \qquad e^{c_J(\theta)} = \int_0^1 e^{\theta^T B_J(x)}\, dx.$$

Thus $p_{J,\theta}$ belongs to a J-dimensional exponential family with sufficient statistics the B-spline functions. Since the B-splines add up to unity, the family is actually of dimension $J-1$ and we can restrict θ to the subset of $\theta \in \mathrm{R}^J$ such that $\theta^T 1 = 0$. The true density p_0 of the observations need not be of the form $p_{J,\theta}$ for some (J, θ).

We assume that the true density p_0 belongs to a Hölder space $C^\alpha[0,1]$, but the order of smoothness α is a-priori only known to belong to a range $\{1, 2, \ldots, \alpha_m\}$, which we take to be integers, for simplicity. We shall construct a prior on the set of log spline densities in two steps. First we construct for each fixed α a prior on the set of log spline densities of a suitable dimension depending on α. Next we combine these priors by putting weights on the different values of α in $\{1, 2, \ldots, \alpha_m\}$.

For given α we need to choose the dimension of the log spline space at least $J_{n,\alpha} = [n^{1/(2\alpha+1)}]$ to achieve good approximation properties for functions belonging to the Hölder space $C^\alpha[0,1]$. We abbreviate the densities $p_{J_{n,\alpha},\theta}$

for $\theta \in \mathrm{R}^{J_{n,\alpha}}$ to $p_{n,\alpha,\theta}$ or even $p_{\alpha,\theta}$. We may also drop the index n in other notation such as $\epsilon_{n,\alpha}$.

It is shown by Stone (1990) that the maximum likelihood estimator $p_{n,\alpha,\hat{\theta}_{n,\alpha}}$ (where $\hat{\theta}_{n,\alpha}$ maximizes the likelihood $\theta \to \prod_{i=1}^n p_{n,\alpha,\theta}(X_i)$ over $\theta \in \mathrm{R}^{J_{n,\alpha}}$ such that $\theta^T 1 = 0$), achieves the rate of convergence $\epsilon_{n,\alpha}$ if the true density p_0 belongs to $C^\alpha[0,1]$, i.e.

$$\int_0^1 (p_{n,\alpha,\hat{\theta}_{n,\alpha}} - p_0)^2(x)\,dx = O_P(\epsilon_{n,\alpha}^2).$$

In Ghosal, Ghosh and van der Vaart (2000) it is shown that a suitable Bayes procedure achieves the same rate. Given a sequence of flat priors $\bar{\Pi}_{n,\alpha}$ on $\mathrm{R}^{J_{n,\alpha}}$ the posterior distributions

$$\Pi_{n,\alpha}(B|X_1,\ldots,X_n) = \frac{\int_{\theta: p_{n,\alpha,\theta}\in B} \prod_{i=1}^n p_{n,\alpha,\theta}(X_i)\,d\bar{\Pi}_{n,\alpha}(\theta)}{\int \prod_{i=1}^n p_{n,\alpha,\theta}(X_i)\,d\bar{\Pi}_{n,\alpha}(\theta)}$$

concentrate almost all of their mass on L_2-balls of radius proportional to $\epsilon_{n,\alpha}$, i.e. for sufficiently large C

$$\Pi_{n,\alpha}\Big(p: \int (p-p_0)^2(x)\,dx \le C\epsilon_{n,\alpha}^2 | X_1,\ldots,X_n\Big) \to 1,$$

almost surely, if the true density p_0 belongs to $C^\alpha[0,1]$. (In both results it is assumed that the true density is also bounded away from zero.)

Both the maximum likelihood estimator and the Bayesian estimator described previously depend on α. They can be made rate-adaptive to α by a variety of means. In the Bayesian set-up, which interests us in this paper, a natural and elegant method to make the procedure rate-adaptive is to put prior weights λ_α on every of the log spline models $\mathcal{P}_{n,\alpha} = \{p_{n,\alpha,\theta} : \theta \in \mathrm{R}^{J_{n,\alpha}}\}$, where $\alpha \in \{1,2,\ldots,\alpha_m\}$, or equivalently put prior weights on the possible degrees of smoothness $\alpha = 1,\ldots,\alpha_m$, and view the previous priors as conditional priors given the model indexed by α. Equivalently, this means using the prior

$$\Pi_n = \sum_\alpha \lambda_\alpha \Pi_{n,\alpha},$$

where $\Pi_{n,\alpha}$ is the image of $\bar{\Pi}_{n,\alpha}$ under the map $\theta \to p_{n,\alpha,\theta}$, on the union $\cup_\alpha \mathcal{P}_{n,\alpha} = \{p_{n,\alpha,\theta} : \theta \in \mathrm{R}^{J_{n,\alpha}}, \alpha = 1,\ldots,\alpha_m\}$ of the models $\mathcal{P}_{n,\alpha}$. (We view $p_{n,\alpha,\theta}$ and $p_{n,\beta,\theta'}$ as different whenever $\alpha \neq \beta$, even if the densities may be the same.) The posterior relative to this prior is

$$\begin{aligned}\Pi_n(B|X_1,\ldots,X_n) &= \frac{\int_B \prod_{i=1}^n p(X_i)\,d\Pi_n(p)}{\int \prod_{i=1}^n p(X_i)\,d\Pi_n(p)} \\ &= \frac{\sum_\alpha \lambda_\alpha \int_{\theta: p_{n,\alpha,\theta}\in B} \prod_{i=1}^n p_{n,\alpha,\theta}(X_i)\,d\bar{\Pi}_{n,\alpha}(\theta)}{\sum_\alpha \lambda_\alpha \int \prod_{i=1}^n p_{n,\alpha,\theta}(X_i)\,d\bar{\Pi}_{n,\alpha}(\theta)}.\end{aligned}$$

Given our finite list of models $\{1, 2, \ldots, \alpha_m\}$ the naive choice of model weights λ_α are the uniform weights $\lambda_\alpha = 1/\alpha_m$. In our asymptotic setting any set of fixed, positive weights λ_α will achieve the same result. It could be benificial to allow the weights to change with n, but we shall not pursue this possibility in this paper.

We assume throughout the paper that the prior measures $\bar{\Pi}_{n,\alpha}$ concentrate inside a (big) block $[-M, M]^{J_{n,\alpha}}$. It is shown in Lemma 5 that restricting θ in $p_{J,\theta}$ to a rectangle $[-M, M]^J$ is equivalent to bounding $p_{J,\theta}$ above and below, the upper and lower bounds approaching infinity and zero if $M \to \infty$. Thus our prior charges only densities that are bounded above and below, and our Bayes procedure cannot be consistent if this is not true for p_0. On the other hand, by making M sufficiently large, we can accomodate any p_0 that is bounded away from zero, no matter how small it may be. With a given, very large M, we can accomodate all "reasonable" p_0.

Furthermore, we assume that the order of the splines involved in the construction of the αth log spline model is at least α. Since we consider only $\alpha \leq \alpha_m$, we can achieve this by choosing all splines to have order $q \geq \alpha_m$.

We assume that the prior distributions $\bar{\Pi}_{n,\alpha}$ are Lebesgue absolutely continuous on the restricted space $\{\theta \in \mathrm{R}^{J_{n,\alpha}} : \theta^T 1 = 0\}$ with densities that vanish outside a big block $[-M, M]^{J_{n,\alpha}}$ and are bounded above and below by $d^{J_{n,\alpha}}$ and $D^{J_{n,\alpha}}$, respectively, for given constants $0 < d \leq D < \infty$. Let $\Pi_{n,1}(\cdot|X_1, \ldots, X_n)$ be the posterior relative to this prior.

THEOREM 1. *If $p_0 \in C^\beta[0,1]$ for some $\beta \in \{1, 2, \ldots, \alpha_m\}$, and p_0 is bounded away from zero and M is sufficiently large, then there exists a constant C such that, in probability under P_0^n,*

$$\Pi_{n,1}\Big(p : \int (p - p_0)^2(x)\, dx \leq C\big(\sqrt{\log n}\,\epsilon_{n,\beta}\big)^2 | X_1, \ldots, X_n\Big) \to 1.$$

The factor $\sqrt{\log n}$ that precedes the rate $\epsilon_{n,\beta}$ in this theorem is disappointing. We would have liked to prove the theorem without this factor, but have not been able to. As can be seen by studying the proof, the logarithmic factor is caused by models $\mathcal{P}_{n,\alpha}$ for $\alpha > \beta$ that approximate the density $p_0 \in C^\beta[0,1]$ within the range $\epsilon_{n,\beta}[1, \sqrt{\log n}]$. If p_0 were really of smoothness level β (and not smoother), then the distance between p_0 and $\mathcal{P}_{n,\alpha}$ is guaranteed to be $(1/J_{n,\alpha})^\beta$ by the approximation properties of splines (see Lemma 4), which is much bigger than $\epsilon_{n,\beta}\sqrt{\log n}$. If the distance of p_0 to $\mathcal{P}_{n,\alpha}$ were really this large, for every n, then the logarithmic factor need not appear, because the posterior would not charge $\mathcal{P}_{n,\alpha}$. However, it appears to be not excluded that the model $\mathcal{P}_{n,\alpha}$ is much closer than the upper bound would predict, infinitely often in n.

We make the observation in the preceding paragraph mathematically precise in the following theorem.

THEOREM 2. *In the preceding theorem the numbers* $\sqrt{\log n}\,\epsilon_{n,\beta}$ *may be replaced by the numbers* $\epsilon^*_{n,\beta}$ *defined by, for a sufficiently large constant* D,

$$\epsilon^*_{n,\beta} = \begin{cases} \sqrt{\log n}\,\epsilon_{n,\beta}, & \text{if } h(p_0, \mathcal{P}_{n,\alpha}) \leq D\sqrt{\log n}\epsilon_{n,\beta}, \text{for some } \alpha > \beta, \\ \epsilon_{n,\beta}, & \text{if } h(p_0, \mathcal{P}_{n,\alpha}) > D\sqrt{\log n}\,\epsilon_{n,\beta}, \text{for all } \alpha > \beta. \end{cases}$$

Thus, for many p_0, the logarithmic factor is unnecessary, as soon as we are in the second case of the definition of $\epsilon^*_{n,\beta}$. We conjecture that to remove the logarithmic factor altogether it is necessary to use more involved priors, for instance based on weights λ_α that change with n, or coefficient-priors $\bar{\Pi}_{n,\theta}$ that are not uniform.

Because we allow only a finite number of models, it is not unreasonable to expect that the posterior probability of the "correct model" converges to one. The posterior probability of model α is given by

$$\frac{\lambda_\alpha \int \prod_{i=1}^n p_{n,\alpha,\theta}(X_i)\, d\bar{\Pi}_{n,\alpha}(\theta)}{\sum_\alpha \lambda_\alpha \int \prod_{i=1}^n p_{n,\alpha,\theta}(X_i)\, d\bar{\Pi}_{n,\alpha}(\theta)}.$$

This expectation of "automatically choosing the right model" may not be justified. A complicating aspect is that the finitely many models in our list are not fixed, but change with n. Furthermore, they may overlap. These same facts also render the proof of the theorem a little delicate.

The remainder of the paper consists of proofs of the theorems. We use the following notation. The symbol $\lesssim$ means "smaller or equal up to a constant times", where the constant is universal or at least fixed within the set-up of the paper (i.e. it may depend on p_0, α_m or M, but not on J or n). We let $\|f\|$ and $\|f\|_\infty$ be the $L_2[0,1]$ and the supremum norm of a function $f : [0,1] \to \mathrm{R}$, and similarly write $\|\theta\|$ and $\|\theta\|_\infty$ for the Euclidean and maximum norm of $\theta \in \mathrm{R}^J$. Because all densities in our set-up are bounded away from zero and infinity, the L_2-norm $\|p - q\|$ between two densities p and q is equivalent to the *Hellinger distance* $\|\sqrt{p} - \sqrt{q}\|$, which we denote by $h(p,q)$. Let P denote the measure corresponding to a density p, and let $Pf = \int f\, dP$.

2. AUXILIARY LEMMAS

In this section we list a number of lemmas that will be used in the proofs of the main theorems.

The usual definition of the space $C^\alpha[0,1]$ (with α an integer) is the set of functions that are $(\alpha-1)$ times differentiable with a Lipschitz $(\alpha-1)$th order derivative. For simplicity, we use the same notation here for the slightly smaller set of functions that are α times continuously differentiable. We let $f^{(\alpha)}$ denote the αth order derivative of f.

LEMMA 3. *Let $q \geq \alpha > 0$. There exists a constant C depending only on q and α such that for every f in $C^\alpha[0,1]$*

$$\inf_{\theta \in \mathrm{R}^J} \left\| \theta^T B_J - f \right\|_\infty \leq C J^{-\alpha} \| f^{(\alpha)} \|_\infty .$$

LEMMA 4. *For any $\theta \in \mathrm{R}^J$,*

$$\begin{aligned} \|\theta\|_\infty \lesssim \|\theta^T B_J\|_\infty \leq \|\theta\|_\infty , \\ \|\theta\| \lesssim \sqrt{J}\, \|\theta^T B_J\| \lesssim \|\theta\| . \end{aligned}$$

LEMMA 5. *For any $\theta \in \mathrm{R}^J$ such that $\theta^T 1 = 0$,*

$$\|\theta\|_\infty \lesssim \| \log p_{J,\theta} \|_\infty \lesssim \|\theta\|_\infty .$$

LEMMA 6. *For every $\theta_1, \theta_2 \in \mathrm{R}^J$ such that $1^T(\theta_1 - \theta_2) = 0$,*

$$\inf_{x,\theta,J} p_{J,\theta}(x) \Big(\frac{\|\theta_1 - \theta_2\|^2}{J} \wedge 1 \Big) \lesssim h^2(p_{J,\theta_1}, p_{J,\theta_2}) \lesssim \sup_{x,\theta,J} p_{J,\theta}(x) \Big(\frac{\|\theta_1 - \theta_2\|^2}{J} \Big),$$

where the infimum and supremum are taken over all θ on the line segment between θ_1 and θ_2 and all $x \in [0,1]$.

The first lemma in this list is the basic approximation lemma for splines and shows that splines of sufficient dimension are well suited to approximating smooth functions. Its proof can be found in De Boor (1978, p170). Note that the order of the splines must be at least the Hölder exponent of the smooth functions. This is fine for our purpose as we seek to adapt only to a maximal degree α_m of smoothness. Lemmas 4-6 are (partly) implicit in Stone (1986, 1990) and can be explicitly found in Ghosal, Ghosh and Van der Vaart (2000). The equivalence of the L_2-norm or infinity-norm on the linear combinations of splines and the Euclidean or maximum norm on the coefficients (up to constants) given by Lemma 4 are consequences of using the B-splines, with their special properties, as a basis.

The following lemma is a consequence of the preceding lemmas.

LEMMA 7. *Let $q \geq \beta$. If $p_0 \in C^\beta[0,1]$ and p_0 is bounded away from zero, then the minimizer $\bar\theta_J$ of $\theta \to \| \log p_{J,\theta} - \log p_0 \|_\infty$ over $\theta \in \mathrm{R}^J$ with $\theta^T 1 = 0$ satisfies*

$$h(p_{J,\bar\theta_J}, p_0) \lesssim \| \log p_{J,\bar\theta_J} - \log p_0 \|_\infty \lesssim J^{-\beta} .$$

B. y Lemma 3 there exists θ_J^* such that

$$\|(\theta_J^*)^T B_J - \log p_0\|_\infty \lesssim J^{-\beta} .$$

In particular, $(\theta_J^*)^T B_J$ is bounded above by a multiple of $\|\log p_0\|_\infty + J^{-\beta}$ and hence is bounded above. Taking exponentials we see that this implies that

$$\|e^{(\theta_J^*)^T B_J} - p_0\|_\infty \lesssim J^{-\beta}.$$

By integrating this inequality, we obtain that $\left|e^{c_J(\theta_J^*)} - 1\right| \lesssim J^{-\beta}$, whence $\left|c_J(\theta_J^*)\right| \lesssim J^{-\beta}$. Because the set of $p_{J,\theta}$ is the same whether θ is restricted to satisfy $\theta^T 1 = 0$ or not, we obtain

$$\begin{aligned}\|\log p_{J,\bar\theta_J} - \log p_0\|_\infty &\le \|\log p_{J,\theta_J^*} - \log p_0\|_\infty \\ &\le \|(\theta_J^*)^T B_J - \log p_0\|_\infty + \left|c_J(\theta_J^*)\right| \lesssim J^{-\beta}.\end{aligned}$$

This proves the second inequality of the lemma. The first inequality follows from the fact that $\log p_0$, and hence $\log p_{J,\bar\theta_J}$ by the preceding display, is uniformly bounded, combined with the inequality $|\sqrt{p} - \sqrt{q}| \le (1/2)e^{M/2}|\log p - \log q|$, uniformly in $p, q \in [0, e^M]$.

The following lemma gives a sufficient condition for the existence of certain tests in terms of the local entropy of a statistical model. For $\epsilon > 0$ and a given metric space (D, d) let $D(\epsilon, \mathcal{P}, d)$ denote the *ϵ-packing number* of D, i.e. the maximal number of points $x_1, \ldots, x_m$ in D such that $d(x_i, x_j) \ge \epsilon$ for $i \ne j$. The lemma is proved in Ghosal, Ghosh and Van der Vaart (2000), following work by Le Cam (1973) and Birgé (1983). The numbers $D(\epsilon)$ in the condition of the following lemma are related to the measures of dimension used by these authors. Up to constants Le Cam (1986) calls the numbers $D(\epsilon)$ the *dimension of $\mathcal{P}$ for the pair (h, ϵ_n)*.

LEMMA 8. *Suppose that for some nonincreasing function $D(\epsilon)$, some $\epsilon_n \ge 0$ and every $\epsilon > \epsilon_n$*

$$\log D\Big(\frac{\epsilon}{2}, \big\{p \in \mathcal{P} : \epsilon \le h(p, p_0) \le 2\epsilon\big\}, h\Big) \le D(\epsilon).$$

Then for every $\epsilon > \epsilon_n$ there exist tests ϕ_n (depending on p_0 and ϵ but not on i) such that, for a universal constant K and every $i \in \mathrm{N}$,

$$\begin{aligned} P_0^n \phi_n &\le e^{D(\epsilon)} e^{-Kn\epsilon^2} \frac{1}{1 - e^{-Kn\epsilon^2}}, \\ \sup_{p \in \mathcal{P} : h(p,p_0) > i\epsilon} P^n(1 - \phi_n) &\le e^{-Kn\epsilon^2 i^2}, \end{aligned}$$

For models $\mathcal{P}$ that are indexed smoothly by a finite-dimensional parameter the number $D(\epsilon)$ in the preceding lemma can be taken equal to the ordinary dimension and hence does not depend on ϵ. This follows from the following lemma, which can be proved by volume arguments. (Cf. Pollard (1990), Lemma 4.1.)

LEMMA 9. *For any norm on* R^J *and any* J, *we have, for* $\eta \geq \epsilon$,

$$\Big(\frac{\eta}{\epsilon}\Big)^J \leq D\Big(\epsilon, \big\{\theta \in \mathrm{R}^J : \|\theta\| \leq \eta\big\}, \|\cdot\|\Big) \leq \Big(\frac{2\eta+\epsilon}{\epsilon}\Big)^J.$$

Next for ease of reference we recall an approximation to the volume of the Euclidean unit ball as the dimension tends to infinity, which is a direct consequence of the explicit formula and Stirling's approximation to the Gamma function.

LEMMA 10. *For* v_J *the volume of the unit ball for the Euclidean norm in* R^J, *as* $J \to \infty$,

$$\sqrt{J}^J v_J = \frac{\sqrt{J}^J \sqrt{\pi}^J}{\Gamma(J/2+1)} = \frac{\sqrt{2\pi e}^J}{\sqrt{\pi J}}(1+o(1)).$$

A final lemma, also taken from Ghosal, Ghosh and Van der Vaart (2000), gives a lower bound on averages of likelihood ratios of product measures.

LEMMA 11. *For every* $\epsilon > 0$ *and probability measure* Π*concentrated on the set* $\big\{p : h^2(p,p_0)\big\|p_0/p\big\|_\infty \leq \epsilon^2\big\}$ *we have, for a universal constant* $B > 0$,

$$P_0^n\Big(\int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi(P) \leq e^{-2n\epsilon^2}\Big) \leq e^{-Bn\epsilon^2}.$$

3. PROOFS OF MAIN RESULTS

We drop the index n in expressions such as $\epsilon_{n,\alpha}$, $p_{n,\alpha,\theta}$, etc. In the proof there will appear constants $A, A_0, A_1, \ldots, B, B_0, \ldots$ that are universal or fixed in our set-up.

Proof of Theorem 1. By Lemma 5 there exists a universal constant A_0 such that $A_0\|\theta\|_\infty \leq \|\log p_{J,\theta}\|_\infty$ for every $\theta \in \mathrm{R}^J$ such that $\theta^T 1 = 0$ and every $J \in \mathrm{N}$. Therefore, for $\bar\theta_J$ as in Lemma 7, we have, for sufficiently large n,

$$\|\bar\theta_{J_\alpha}\|_\infty \leq \frac{1}{A_0}\|\log p_{\alpha,\bar\theta_{J_\alpha}} - \log p_0\|_\infty + \frac{1}{A_0}\|\log p_0\|_\infty \leq A_1 J_\alpha^{-\beta} + (1/2)M \leq M,$$

eventually, if $\|\log p_0\|_\infty \leq (1/2)A_0 M$. We shall assume that p_0 and M are matched to each other in this way. We also assume that $M \geq 1$.

Given the value of M that is now fixed, the numbers $\|\log p_{\alpha,\theta}\|_\infty$ are uniformly bounded in α and in $\theta \in \mathrm{R}^{J_\alpha}$ such that $\|\theta\|_\infty \leq M$, by Lemma 5. In other words $p_{\alpha,\theta}$ is bounded away from zero and infinity, uniformly in θ and α

with $\|\theta\|_\infty \le M$. For simplicity of notation we redefine $\mathcal{P}_\alpha$ as $\{p_{\theta,\alpha} : \theta \in \mathrm{R}^{J_\alpha}, \|\theta\|_\infty \le M\}$.

Let θ_α minimize the Hellinger distance $\theta \mapsto h(p_{\alpha,\theta}, p_0)$ over $\theta \in \mathrm{R}^{J_\alpha}$ such that $\theta^T 1 = 0$ and $\|\theta\|_\infty \le M$. Thus $h(p_{\alpha,\theta_\alpha}, p_0)$ is the Hellinger distance of the true density to the model $\mathcal{P}_\alpha$ consisting of all densities $p_{\theta,\alpha}$. Because $\|\bar\theta_{J_\alpha}\|_\infty \le M$, we can conclude from Lemma 7 that $h(p_{\alpha,\theta_\alpha}, p_0) \le h(p_{\alpha,\bar\theta_{J_\alpha}}, p_0) \lesssim J_\alpha^{-\beta}$.

For every $\epsilon > 0$ we have

$$\big\{p_{\alpha,\theta} : h(p_{\alpha,\theta}, p_0) \le \epsilon\big\} \subset \big\{p_{\alpha,\theta} : h(p_{\alpha,\theta}, p_{\alpha,\theta_\alpha}) \le 2\epsilon\big\}.$$

Indeed, either $h(p_{\alpha,\theta_\alpha}, p_0) > \epsilon$ and hence the set on the left is empty, or $h(p_{\alpha,\theta_\alpha}, p_0) \le \epsilon$ and then the inclusion follows by the triangle inequality. Combining this with Lemma 6 we see that there exists a constant A such that

$$\big\{p_{\alpha,\theta} : h(p_{\alpha,\theta}, p_0) \le \epsilon, \|\theta\|_\infty \le M\big\} \subset \Big\{p_{\alpha,\theta} : \frac{\|\theta - \theta_\alpha\|}{\sqrt{J_\alpha}} \wedge 1 \le \epsilon(1/2)A\Big\}.$$

Because $\|\theta - \theta_\alpha\| \le \sqrt{J_\alpha}\|\theta - \theta_\alpha\|_\infty \le 2M\sqrt{J_\alpha}$ for every θ with $\|\theta\|_\infty \le M$ we have the further inclusion

$$\big\{p_{\alpha,\theta} : h(p_{\alpha,\theta}, p_0) \le \epsilon, \|\theta\|_\infty \le M\big\} \subset \Big\{p_{\alpha,\theta} : \|\theta - \theta_\alpha\| \le A\sqrt{J_\alpha}\epsilon M\Big\}, \quad (1)$$

because $\{x \ge 0 : x \wedge 1 \le \epsilon, x \le M\} \subset \{x : x \le \epsilon M\}$ for every $\epsilon > 0$ and $M \ge 1$.

In view of this inclusion and Lemma 6 there exist constants B_i such that, for every $\epsilon > 0$,

$$\begin{aligned} &D\Big(\frac{\epsilon}{2}, \big\{p_{\alpha,\theta} : h(p_{\alpha,\theta}, p_0) \le 2\epsilon, \|\theta\|_\infty \le M\big\}, h\Big) \\ &\quad\le D\Big(B_0\epsilon\sqrt{J_\alpha}, \big\{\theta \in \mathrm{R}^{J_\alpha} : \|\theta - \theta_\alpha\| \le A\sqrt{J_\alpha}2\epsilon M\big\}, \|\cdot\|\Big) \\ &\quad\le \Big(\frac{B_1 A\sqrt{J_\alpha}\epsilon M}{\epsilon\sqrt{J_\alpha}}\Big)^{J_\alpha} = e^{J_\alpha B}, \end{aligned}$$

where the last inequality follows from Lemma 9.

Let $\bar\epsilon_\alpha = \epsilon_\alpha \vee \epsilon_\beta$. Because $J_\alpha = n\epsilon_\alpha^2 \le n\bar\epsilon_\alpha^2$, we may apply Lemma 8 with $\epsilon = E\bar\epsilon_\alpha$ for a large constant E and $D(\epsilon) = e^{Bn\bar\epsilon_\alpha^2}$ to obtain tests ϕ_α (not depending on $i \in \mathrm{N}$) such that for every $i \in \mathrm{N}$

$$P_0^n\phi_\alpha \lesssim e^{(B - KE^2)n\bar\epsilon_\alpha^2}, \quad (2)$$

$$\sup_{p \in \mathcal{P}_\alpha : h(p,p_0) > iE\bar\epsilon_\alpha} P^n(1 - \phi_\alpha) \le e^{-KE^2 i^2 n\bar\epsilon_\alpha^2}. \quad (3)$$

By (2), for every α, if E is large enough to ensure that $B - KE^2 < 0$,

$$\mathrm{E}_0 \Pi_n \Big(p : h(p, p_0) > C\epsilon_{n,\beta} | X_1, \ldots, X_n \Big) \phi_\alpha \leq P_0^n \phi_\alpha \to 0.$$

Set $\epsilon_{\alpha,\beta} = \epsilon_\beta \vee h(p_0, p_{\alpha,\theta_\alpha})$. By the triangle inequality,

$$\begin{aligned} \big\{ p \in \mathcal{P}_\alpha : h(p, p_{\alpha,\theta_\alpha}) \leq \epsilon_{\alpha,\beta} \big\} &\subset \big\{ p \in \mathcal{P}_\alpha : h(p, p_0) \leq 2\epsilon_{\alpha,\beta} \big\} \\ &\subset \Big\{ p \in \mathcal{P}_\alpha : h^2(p, p_0) \Big\| \frac{p_0}{p} \Big\|_\infty \leq F\epsilon_{\alpha,\beta}^2 \Big\}, \end{aligned}$$

for some constant F, because the quotients p_0/p are uniformly bounded. By Lemma 11 there exist events $A_{n,\alpha}$ with $P_0^n(A_{n,\alpha}) \geq 1 - \exp(-F_2 n \epsilon_{\alpha,\beta}^2)$ on which

$$\begin{aligned} \int \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi_\alpha(p) &\geq \int_{p : h(p, p_{\alpha,\theta_\alpha}) \leq \sqrt{F}\epsilon_{\alpha,\beta}} \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi_\alpha(p) \\ &\geq e^{-2Fn\epsilon_{\alpha,\beta}^2} \Pi_\alpha \big(p : h(p, p_{\alpha,\theta_\alpha}) \leq \sqrt{F}\epsilon_{\alpha,\beta} \big) \\ &\geq e^{-2Fn\epsilon_{\alpha,\beta}^2} \inf_\theta \bar{\pi}_\alpha(\theta) (G\epsilon_{\alpha,\beta})^{J_\alpha} \sqrt{J_\alpha}^{J_\alpha} v_\alpha, \end{aligned}$$

where v_α is the volume of the Euclidean unit ball in R^{J_α}. In the last step we use the inclusion

$$\big\{ \theta \in \mathrm{R}^{J_\alpha} : \|\theta - \theta_\alpha\| \leq \epsilon_{\alpha,\beta} 2G\sqrt{J_\alpha} \big\} \subset \big\{ \theta \in \mathrm{R}^{J_\alpha} : h(p_{\alpha,\theta}, p_{\alpha,\theta_\alpha}) \leq \sqrt{F}\epsilon_{\alpha,\beta} \big\},$$

which follows from Lemma 6. We have not shown that the ball on the left side is also contained in the set $\big\{ \theta \in \mathrm{R}^{J_\alpha} : \|\theta\|_\infty \leq M \big\}$, but because $\|\theta_\alpha\|_\infty \leq M$, we must have that at least $(1/2)^{J_\alpha}$ of its volume intersects $\big\{ \theta \in \mathrm{R}^{J_\alpha} : \|\theta\|_\infty \leq M \big\}$.

Define $B_\alpha(\epsilon) = \big\{ p \in \mathcal{P}_\alpha : h(p, p_0) \leq \epsilon \big\}$ and, for $i \in \mathrm{N}$,

$$S_{\alpha,i} = \big\{ p \in \mathcal{P}_\alpha : iE\bar{\epsilon}_\alpha < h(p, p_0) \leq (i+1)E\bar{\epsilon}_\alpha \big\}.$$

Because $\bar{\epsilon}_\alpha = \epsilon_\beta$ for $\alpha \geq \beta$ and $\bar{\epsilon}_\alpha = \epsilon_\alpha > \epsilon_\beta$ for $\alpha < \beta$, we have

$$\begin{aligned} &\big\{ p : h(p, p_0) > CE\epsilon_\beta \big\} \\ &\quad \subset \cup_\alpha \cup_{i \geq C} S_{\alpha,i} \cup \cup_{\alpha < \beta} \big\{ p \in \mathcal{P}_\alpha : CE\epsilon_\beta < h(p, p_0) \leq CE\epsilon_\alpha \big\} \\ &\quad \subset \cup_\alpha \cup_{i \geq C} S_{\alpha,i} \cup \cup_{\alpha < \beta} B_\alpha(CE\epsilon_\alpha). \end{aligned}$$

By Fubini's theorem and (3)

$$\mathrm{E}_0 \int_{S_{\alpha,i}} \prod_{i=1}^n \frac{p}{p_0}(X_i)(1 - \phi_\alpha)\, d\Pi_\alpha(p) \leq e^{-KE^2 i^2 n \bar{\epsilon}_\alpha^2} \Pi_\alpha(S_{\alpha,i}).$$

By Fubini's theorem and the inequality $P_0(p/p_0) \le 1$, we have for every set C,

$$\mathrm{E}_0 \int_C \prod_{i=1}^n \frac{p}{p_0}(X_i)\, d\Pi_\alpha(p) \le \Pi_\alpha(C).$$

Because $n\epsilon_{n,\beta}^2 \to \infty$ we have $P_0^n(A_{n,\alpha}) \to 1$ for each α, whence, because they are only finitely many α, the events $A_{n,\alpha}$ satisfy $P_0^n(\cap_\alpha A_{n,\alpha}) \to 1$. Furthermore,

$$\begin{aligned}
&\mathrm{E}_0 \Pi_n\big(h(p,p_0) > CE\epsilon_\beta | X_1, \ldots, X_n\big)(1 - \max_\alpha \phi_{n,\alpha}) 1_{\cap_\alpha A_{n,\alpha}} \\
&\quad \le \frac{\mathrm{E}_0 \sum_\alpha \sum_{i \ge C} \int_{S_{\alpha,i}} \prod_i \frac{p}{p_0}(X_i)(1-\phi_\alpha)\, d\Pi_\alpha(p)}{\sum_\alpha \lambda_\alpha \int \prod_i \frac{p}{p_0}(X_i)\, d\Pi_\alpha(p)} \\
&\qquad + \frac{\mathrm{E}_0 \sum_{\alpha<\beta} \int_{B_\alpha(CE\epsilon_\alpha)} \prod_i \frac{p}{p_0}(X_i)\, d\Pi_\alpha(p)}{\sum_\alpha \lambda_\alpha \int \prod_i \frac{p}{p_0}(X_i)\, d\Pi_\alpha(p)} \\
&\quad \le \sum_{\alpha \ge \beta} \frac{\sum_{i \ge C} e^{-KE^2 i^2 n\epsilon_\beta^2} \Pi_\alpha\big(B_\alpha((i+1)E\epsilon_\beta)\big)}{\lambda_\alpha e^{-2Fn\epsilon_{\alpha,\beta}^2} (dG\epsilon_{\alpha,\beta}\sqrt{J_\alpha})^{J_\alpha} v_\alpha + \lambda_\beta e^{-2Fn\epsilon_{\beta,\beta}^2} (dG\epsilon_{\beta,\beta}\sqrt{J_\beta})^{J_\beta} v_\beta} \\
&\qquad + \sum_{\alpha<\beta} \frac{\sum_{i \ge C} e^{-KE^2 i^2 n\epsilon_\alpha^2} \Pi_\alpha\big(B_\alpha((i+1)E\epsilon_\alpha)\big)}{\lambda_\beta e^{-2Fn\epsilon_{\beta,\beta}^2} (dG\epsilon_{\beta,\beta}\sqrt{J_\beta})^{J_\beta} v_\beta} \\
&\qquad + \sum_{\alpha<\beta} \frac{\Pi_\alpha\big(B_\alpha(CE\epsilon_\alpha)\big)}{\lambda_\beta e^{-2Fn\epsilon_{\beta,\beta}^2} (dG\epsilon_{\beta,\beta}\sqrt{J_\beta})^{J_\beta} v_\beta}.
\end{aligned}$$

The proof of Theorem 1 is complete once it is shown that every of the three sums on the far right converges to zero as $C = C_n = \sqrt{\log n}$. For each of the three sums it suffices to consider one α-term at a time, because the set of possible α is finite. In view of (1)

$$\Pi_\alpha\big(B_\alpha(\epsilon)\big) \le D^{J_\alpha} \big(A\sqrt{J_\alpha}\epsilon M\big)^{J_\alpha} v_\alpha.$$

Furthermore, by the definition of θ_α and Lemma 7, we have the inequalities $h(p_0, p_{\alpha,\theta_\alpha}) \le h(p_0, p_{\alpha,\bar\theta_{J_\alpha}}) \lesssim J_\alpha^{-\beta}$. In particular, $h(p_0, p_{\beta,\bar\theta_{J_\beta}}) \lesssim J_\beta^{-\beta} = \epsilon_\beta$, so that $\epsilon_\beta \le \epsilon_{\beta,\beta} \le H\epsilon_\beta$.

A typical α-term of the first sum (with $\alpha \ge \beta$) is bounded by

$$\sum_{i \ge C} \frac{e^{-KE^2 i^2 n\epsilon_\beta^2} \big(DA\sqrt{J_\alpha}(i+1)E\epsilon_\beta M\big)^{J_\alpha} v_\alpha}{\lambda_\beta e^{-2F_1 n\epsilon_\beta^2} (dG_1\epsilon_\beta\sqrt{J_\beta})^{J_\beta} v_\beta}.$$

This converges to zero for $C = C_n = \sqrt{\log n}$ and sufficiently large E, in view of Lemma 10.

Next, consider a typical α-term of the second sum (with $\alpha < \beta$). For $\alpha < \beta$ we have $\epsilon_\alpha > \epsilon_\beta$ and $J_\alpha > J_\beta$. A typical term is bounded above by

$$\sum_{i \geq C} \frac{e^{-KE^2 i^2 n\epsilon_\alpha^2} 1}{\lambda_\beta e^{-2FHn\epsilon_\beta^2} (dG\epsilon_\beta \sqrt{J_\beta})^{J_\beta} v_\beta}.$$

This converges to zero for fixed C in view of Lemma 10, as $n\epsilon_\alpha^2 = J_\alpha \gg J_\beta \log(1/\epsilon_\beta) \gg J_\beta = n\epsilon_\beta^2$.

Finally, a typical α-term of the third sum (with $\alpha < \beta$) is bounded above by

$$\frac{(DA\sqrt{J_\alpha} CE\epsilon_\alpha M)^{J_\alpha} v_\alpha}{\lambda_\beta e^{-2FHn\epsilon_\beta^2} (dG\epsilon_\beta \sqrt{J_\beta})^{J_\beta} v_\beta}.$$

This converges to zero for fixed C at the order $e^{-LJ_\alpha \log n}$ for some constant L.

Proof of Theorem 2. In the conclusion of the proof of Theorem 1 it was shown that three terms converge to zero. The second and third terms converge to zero for a fixed value of the constant C, whereas the first term converges to zero as $C = C_n$ converges to zero at logarithmic rate. We only need to change the arguments for the first term. The term given by $\alpha = \beta$ does not cause problems. If $\alpha > \beta$ and $h(p_0, \mathcal{P}_\alpha) \geq D\sqrt{\log n}\epsilon_\beta$ (which is necessary for $\epsilon_{n,\beta}^* \neq \epsilon_{n,\beta}$), then the balls $B_\alpha\big((i+1)E\epsilon_\beta\big)$ are empty for $i \lesssim \sqrt{\log n}$ and hence their prior masses are zero. Thus the terms of the sum for $C \leq i \lesssim \sqrt{\log n}$ vanish, and it is not a loss of generality to assume $C = C_n = \sqrt{\log n}$.

REFERENCES

1. Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Zeitschrift Wahrscheinlichkeitstheorie* **65**. 181–238.
2. de Boor, C de (1978). A Practical Guide to Splines. Springer, New York.
3. Ghosal, S., Ghosh, J.K., van der Vaart, A.W. (2000). Convergence rates of posterior distributions. *Annals of Statistics***28**. 500–531.
4. Kolmogorov, A.N., Tikhomirov, V.M. (1961). Epsilon-entropy and epsilon-capacity of sets in function spaces *American Mathematical Society Translations, series 2***17**. 277–364.
5. Le Cam, L.M. (1973). Convergence of estimates under dimensionality restrictions. *Annals of Statistics***2**. 38–53.
6. Le Cam, L.M. (1986). Asymptotic Methods in Statistical Decision Theory. Springer, New York.
7. Pollard, D. (1990). Empirical Processes: Theory and Applications. Society for Industrial and Applied Mathematics, Philadelphia. NSF-CBMS Regional Conference Series in Probability and Statistics **2**. Institute of Mathematical Statistics and American Statistical Association

8. Schumaker, L.L. (1981). Spline functions basic theory. Wiley and Sons.

9. Stone, C.J. (1986). The dimensionality reduction principle for generalized additive models. *Annals of Statistics***14**. 590–606.

10. Stone, C.J. (1990). Large-sample inference for log-spline models *Annals of Statistics***18** 717–741.